INF 2178 – Technical Assignment 4

1. **Introduction & Background of Data**

   The assignment dataset is derived from a subset of data from a longitudinal study on MRI results of patients with/without dementia will be used to explore within-subject design and conduct statistical power test. The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. We delve into dataset named 'INF2178_A4_data.csv' (accessible in the GitHub repository), tracking how the patient's nWBV changes across multiple aspects.

2. **Data Cleaning and Processing**

The raw dataset has a total of **16 columns** with **294** entities **(rows)** and **two IDs variables** – **Subject ID** and **MRI ID**. After the initial review of the dataset, we thought a basic data cleaning was necessarily for the scope of our analysis. Below we showed the observations of our dataset.

A. Observations and Considerations:

Since our analysis is quantitative, we must work on the specific columns from the raw dataset. We could ignore the first column (Unnamed: 0) as it is not necessary. We have a total **7 Numeric Variables** and **6 Categorical Variables**.

B. The raw dataset has **15 missing values in SES** and **1 missing value in MMSE**. And we should drop the missing values to make sure data accuracy and integrity.

3. **Exploratory Data Analysis (EDA)**

After conducted data cleaning to our new working dataset, we created with a comprehensive EDA to examine insights that could potentially lead to explore our research questions.

Research Question #1: Is the relationship between the different groups and their socioeconomic status (SES)?

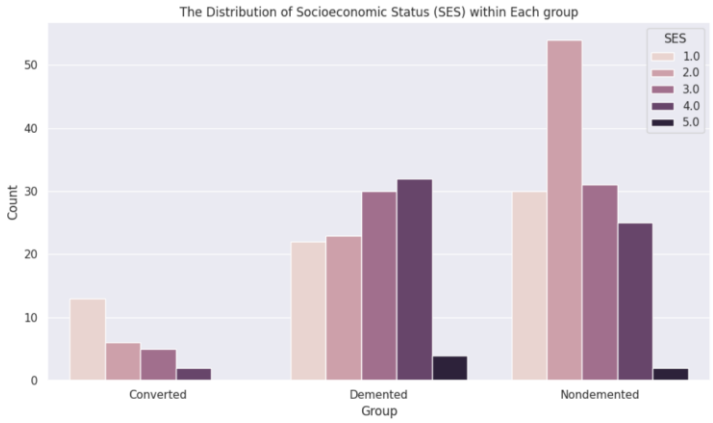

Figure 1: Bar-chart of Distribution of SES by Different Groups

As we delve into how the socioeconomic status have impact the individuals with and without dementia. Referring to *Figure 1*, the bar-chart demonstrates a noticeable sign that individuals with higher socioeconomic status will be less likely to get demented. This may indicate that people' socioeconomic status have a relationship with the

occurrence of dementia, especially it showed a higher proportion of individuals with SES level at 3 and 4 within the demented group. It can be inferred that there is a tendency of getting dementia towards lower socioeconomic status. In addition, the low proportion of individuals with SES at level 5 examines a potential problem whereas the lower-income populations lack of sufficient medical resources for examination and treatment. Therefore, policymakers should increase medical resources or aids so that they can get access to necessary health checks and treatment services.

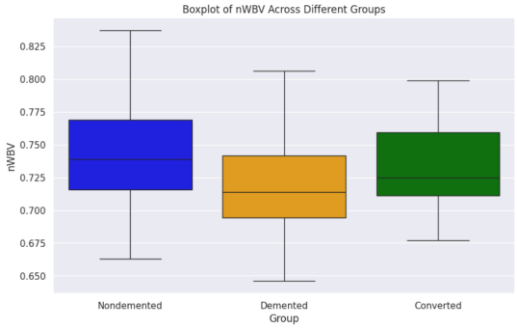

Figure 2: Boxplot of nWBV by Different Groups

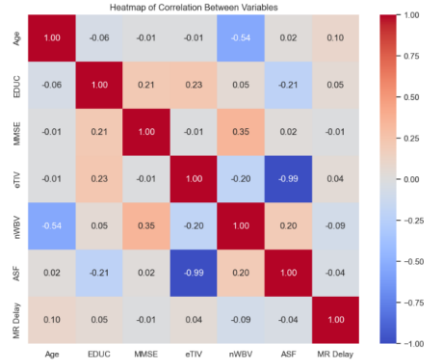

Figure 3: Correlation Matrix of Quantitative Data

Based on the boxplot above, we can observe the changes by groups, offering a comparison of nWBV across Demented, Nondemented and Converted groups. While nondemented individuals reveal a higher median nWBV and less variability (narrower IQR), those with dementia or converted show lower median values and greater variability.

Referring *Figure 3* to reveal a holistic overview of the relationships between variables. As seen on the correlation matrix, nWBV have a moderately negative correlated with age, in which maybe relevance to researchers on studying neurodegenerative diseases or aging process of the brain. Other correlations are relatively weak, demonstrating that nWBV is an independent of education level, estimated total intracranial volume (eTIV), atlas scaling factor (ASF), and MRI Delay.

4. nWBV Changes over Visits and Groups

Research Question #2: How does the normalized whole brain volume (nWBV) changes over visits in individuals with and without dementia, and is there an interaction effect between the visit number and dementia status on nWBV changes?

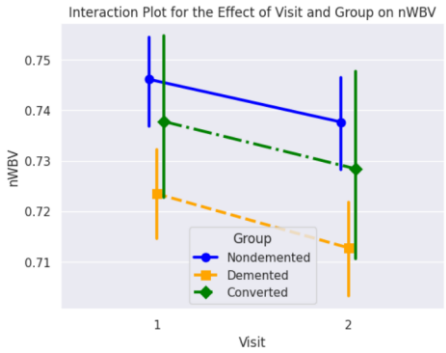

Figure 4: Interaction Plot of Showing Effect of Visit and Group on nWBV

Source	SS	DF1	DF2	MS	F	P-value	np2
Group	0.033	2	134	0.017	6.384	0.002	0.087
Visit	0.006	1	134	0.006	89.376	0.001	0.4
Interaction	0	2	134	0	1.63	0.2	0.024

Table 1: ANOVA Table Summary Results

Referring to *Figure 4*, the interaction plot shows that nWBV decreases over time in all groups, with the most significant change in the converted group, which surpasses the demented group by the second visit. This indicates not only an overall trend of declining brain volume associated with dementia but also highlights a potentially greater rate of decline in those who have converted to dementia between the visits.

Based on the ANOVA table summary presented, the significant differences were observed in nWBV both between different groups (p-value=0.002) and across various visits or time points (p-value<0.001). These findings reveal that both the between-groups factor 'Group' and the within-subjects factor 'Visit' independently influence nWBV. However, the interaction between these two factors did not show a significant effect on nWBV (p-value=0.2), inferring that the effect of one factor does not depend on the level of the other.

Contrast	Visit	A	B	P-value
Visit	-	1	2	0.001
Group	-	Converted	Demented	0.171
Group	-	Converted	Nondemented	0.527
Group	-	Demented	Nondemented	0.001
Visit * Group	1	Converted	Demented	0.162
Visit * Group	1	Converted	Nondemented	0.648
Visit * Group	1	Demented	Nondemented	0.001
Visit * Group	2	Converted	Demented	0.19
Visit * Group	2	Converted	Nondemented	0.429
Visit * Group	2	Demented	Nondemented	0.001

Table 2: Post-Hoc Test Results for ANOVA

Furthermore, we conducted the post-hoc analysis displays statistically significant differences in nWBV over time and between Demented and Nondemented groups (as p-value<0.001), with a specific time-dependent interaction effect notable only between

these two groups.

Assumption Checks for Running Mixed-Effect ANOVA:

Assumption 1: Normality of Residuals

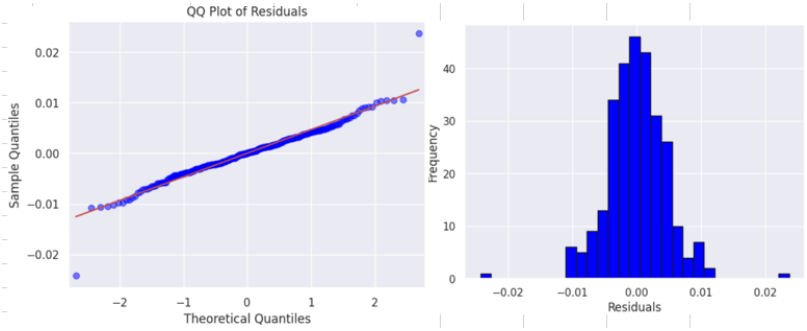

Figure 5:QQ plot and Histogram for Residuals

As we can see, the Q-Q plot and histogram are used to check the assumption of normality. The histogram is **roughly a normal distribution**. The Q-Q plot helps this analysis, the most data points lie along the line, which indicates that the **residuals are normal distributed**, except two potential outliers or extreme values.

Assumption 2: Homogeneity of Variances (Levene's test)

Parameter	Value
Test Statistic (W)	0.489
p-value	0.614

Table 2: Levene's Test Results

Based on Levene's test results in table 2, is crucial for checking the assumption of homogeneity of variances in the ANOVA analysis. Since the p-value is higher than significance level, which indicates that the **assumption of equal variances is satisfied**.

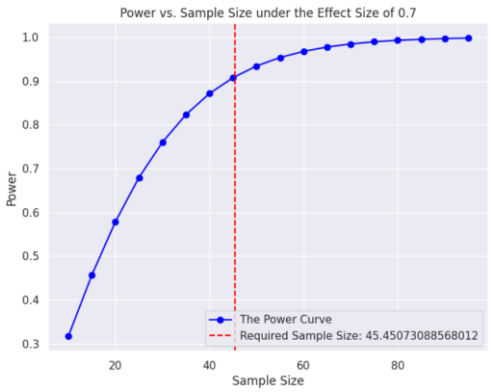

Figure 6: Power Plot for T-test

In conclusion, we used the given effect size (0.7), statistical power (0.91), and alpha level (0.05) to calculate the appropriate **sample size** for each group should be a total **46 individuals** as required. This is an essential step for theoretical experimental design, ensuring the study have enough sample size to detect the effects of interest. As we can observe from the power plot (*Figure 6*), there is a positive correlation between the sample size and the statistical power of the test: as the desired power increases, so does the necessary sample size. This relationship is crucial to make sure the credibility of the experiment's results. Higher statistical power means there is a greater possibility of correctly detecting a true effect.